Automatic detection of short-term sleepiness state.
Sequence-to-Sequence modelling with global attention mechanism.

Edward L. Campbell¹, Laura Doctó-Fernández¹, Carmen García-Mateo¹, Andre Wittenborn², Jarek Krajewski², Nicholas Cummins³

¹GTM research group, AtlantTTic Research Center, University of Vigo, Spain
²Engineering Psychology, Rhenish University of Applied Science, Cologne, Germany
³Institute of Psychiatry, Psychology and Neuroscience, King’s College London, London, UK

Abstract
The Continuous Sleepiness detection task was a Sub-Challenge developed in the 2019 INTERSPEECH Computational Paralinguistics Challenge (ComParE). The associated speech corpus has been a reference in last years for the speech-based detection of sleepiness conditions. In this paper, we proposed a Sequence-to-Sequence model with global attention mechanism to accomplish this detection task. To the best of the authors’ knowledge, this is the first such an approach has been proposed for this task. Given the smaller size of this corpus, we utilise a small batch size, and augment our system with a score ensembling strategy to deliver the final decision. Despite the high complexity of our approach, it produces exceptionally competitive performances on the test-set, producing the second best performance to date. This result highlights the benefits of using deep-learning approaches, even with smaller sized speech-corpora.

Index Terms: sleepiness detection, 2019 Interspeech Computational Paralinguistics Challenge (ComParE), Sequence-to-Sequence modelling, Global attention mechanism.

1. Introduction
The feeling of being tired and wanting to sleep (sleepiness) can affect our every-day-life if it occurs in a frequently basis [1]. Sleep disorders may be accompanied by a variety of daytime complaints and symptoms, including fatigue, decreased energy, and mood disturbances [1]. Symptoms of anxiety or depression that do not meet criteria for a specific mental disorder may be present, as well as an excessive focus on the perceived effects of sleep loss on daytime functioning [1]. The detection and estimation of short-term subjective sleepiness on healthy subjects has applications in the monitoring of performances for tasks requiring high cognitive loads, such as driving, steering, reading and learning [2, 3, 4, 5, 6].

Regarding the automatic sleepiness state detection, there are many different patterns coming from subjects’ speech [7] that make this process feasible. Sleepy speech, when compared with rested speech may exhibit acoustic changes in prosody, articulation and speech quality [8]. In prosody, sleepy speech can feature monotonic and flattened intonation, shifted speech rate, or reduced syllable duration due to slowed cognitive speech planning [8]. Articulation issues such as less crisp pronunciation, mispronunciations, abrupt articulatory changes, slurred, speech errors, or hesitations can appear due to, e.g., impaired motor coordination processes and aversion of spending compensatory effort [9, 8]. In the case of speech quality, it might be affected by tensed, nasal, or breathy speech due to, e.g., impaired coordination of velum closure [10, 8].

The Interspeech 2011 Speaker State Challenge was the first of its kind in focusing on the detection of sleepiness condition from speech [11]. It was approached as a binary classification task on the Sleepy Language Corpus, with 21 hours of speech recordings and 99 subjects. The speech data consisted of different tasks: isolated vowels [1], read speech, commands/requests, four simulated pilot-air traffic controller communication statements, a description of a picture and a regular lecture [11]. In that occasion, Support Vector Machine (SVM) was the predominant selected classifier for performing the detection task on acoustic and prosodic features, achieving an unweighted accuracy of 70.3% as the best baseline system [11].

Eight years later, the Interspeech 2019 (IS19) challenge on continuous sleepiness estimation introduced the SLEEP corpus (also referred to as the Dusseldorf Sleepy Language Corpus) [12]. The SLEEP corpus, with more than 16,000 samples, seemed more suited for regression (that requires more data) than its previous edition, which contains 9,000 samples. Although all these efforts, there still exist pitfalls and problems in analyzing sleepiness through voice [2]. Few deep learning systems have been proposed through the years [13, 14]. For example, the benefits of Sequence-to-Sequence (S2S) models have not been investigated yet. These models, in combination with attention mechanisms, have shown outstanding results on mental health tasks such as the detection of Major Depressive Disorders [15, 16, 17].

Therefore, this paper aims to validate the efficacy of S2S model as a long-sequence processor in the detection of short-term sleepiness condition, with the SLEEP corpus as experimental framework. A group of acoustic and prosodic features are presented as well, with the goal of keeping a tiny but discriminative feature space that help us to identify patterns relative to the sleepiness state.

The remainder of this paper is organized as follows: Section 2 describes the experimental corpus. Our proposed S2S model architecture is depicted in Section 3. Then, results are discussed in Section 4, and conclusions are given in Section 5.

2. SLEEP corpus
The Continuous Sleepiness Estimation corpus, as a subset of the SLEEP (Dusseldorf Sleepy Language) Corpus, was created
at the Institute of Psychophysiology, Düsseldorf, and the Institute of Safety Technology, University of Wuppertal, Germany [12]. There was a total of 915 subjects, from which 364 were females and 551 males. Participants’ age are between 12 and 84 years old, with an average of 27.6 and standard-deviation of 11.

Regarding recording conditions, interviewees were in quiet rooms using a microphone/headset/hardware setup with the tasks to perform presented on a computer in front of them. Recording frequency sampling was 44.1 kHz and down-sampled to 16 kHz, with a quantisation of 16 bit. Participants were asked to follow different reading passages and speaking tasks, e.g., the description of their last weekend.

To measure the subjective sleepiness level of participants, they had to report on the well-established Karolinska Sleepiness Scale (KSS) [18]. Subjects had to indicate which level best reflects the psycho-physical state experienced in the last 10 minutes. Therefore, it represents a measure of situational sleepiness sensitive to fluctuations [18]. The KSS scoring scale is as follows:

- 1: Extremely alert
- 2: Very alert
- 3: Alert
- 4: Rather alert
- 5: Neither alert nor sleepy
- 6: Some signs of sleepiness
- 7: Sleepy, but no difficulty remaining awake
- 8: Sleepy, some effort to keep alert
- 9: Extremely sleepy, fighting sleep

Figure 1 illustrates the score distribution in the training set, which is quite similar in the development and testing set as well. Only a small percent of participants indicate extreme conditions close to alert or extremely sleepy.

<table>
<thead>
<tr>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td># Samples</td>
<td>5564</td>
<td>5328</td>
</tr>
</tbody>
</table>

Figure 2 illustrates the interview duration per set. Boxplot diagram

The next section describes our proposed system for the automatic detection of sleepiness condition.

### 3. Sequence-to-Sequence model for sleepiness state classification.

The detection process has two main steps: feature extraction and classification. As speech’s features, we extracted a group of acoustic, prosody and articulatory features compound by 13 Mel-Spectrum coefficients with their first and second derivatives, the zero-crossing rate, the energy signal and the spectral centroid [19]. It makes a feature vector of 42 coefficients. Our selection was based on the idea of highlighting acoustic speech characteristics that appeared for sleepy subjects in previous studies and were described in Section 1 [10, 8, 9]. Moreover, these features have delivered good results in the task of Major Depressive Disorders detection [20, 21, 22, 23].

A Sequence-to-Sequence (S2S) model (Figure 3) with global attention mechanism [24, 25] was used as classifier to process the long-term sequence delivered by our feature vector. This classifier is based on our previous work at [15]. However, we used a smaller window context (0.5 seconds) to frame the feature sequences because of the small duration of participant’s interviews in the Sleepiness corpus.

S2S is built on Recurrent Neural Network (RNN) architectures which have been shown to be effective in the processing of sequential data [26]. RNNs ability to track and store dependencies throughout a sequence has been key in tasks such as Stock Price Pattern Recognition [27, 28] and health care [29]. The use of RNN’s is further justified given that depression has been shown to alter temporal properties of speech; [30, 31].

S2S is a modelling paradigm that uses two sets of RNNs to convert one sequence of items in one domain into a sequence in
Figure 3: Sequence to sequence model architecture with global attention mechanism

another domain [32]. The first RNN network is known as the encoder and the second one as decoder. The encoder learns to processes each item of an input sequence and converts this information into a fixed (static) representation vector know as the context vector. The decoder then learns to converts this static representation into new sequence.

The core component of the encoder and decoder set-up are the RNN blocks. They are developed by Gated Recurrent Unit (GRU) layers with a hidden size of 64 (Fig. 3). In the encoder’s input, a Batch Normalization layer is applied for decreasing the training time of the model. Linear transformations are also applied for guarantee matrix compatibility between some consecutive blocks. In the decoder, we added a global attention mechanism [25] in order to consider information relevance when processing the output.

To sum up, the methodology of this system is the following: an interview is divided into a sequence of features vectors which are classified by a Sequence-to-Sequence model with a global attention mechanism. Each sequence gets a score assigned by the classifier. Finally, those sequence scores are averaged to get a final label that represents the participant sleepiness state.

4. Results and Discussion

The Spearman’s correlation coefficient (\(\rho\)) [33] was the metric proposed in The INTERSPEECH 2019 Computational Paralinguistics Challenge to evaluate systems performance. Therefore, we also use this measure so we can compare our system performance with other published results.

The random initialization of a Deep Neural Network’s weights can influence its performance. Furthermore, we trained 10 different models initialization and evaluated them on the development set, achieving a mean \(\rho\) of 0.243 and a standard-deviation of 0.015. This shows the performance our architecture is quite stable through different runs.

However to assist in countering the smaller size of the training data, an score ensemble strategy was implemented as well. Here, the score of the previously trained systems were averaged to have a final decision score. We tried clustering the previous systems in a group of five and another of ten. In each case, we averaged their scores to get a final detection decision. Both of them showed similar results, outperforming the 10 stand-alone systems. Table 2 illustrates the average performance of the 10 stand-alone systems and the score ensembling strategy using 10 systems.

Table 2: Results on the development set. Spearman’s correlation coefficient of the ensembling system and the average performance of the stand-alone one.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2S</td>
<td>.243</td>
</tr>
<tr>
<td>S2S-ensembling</td>
<td>.257</td>
</tr>
</tbody>
</table>

A batch size of 128 was selected to decrease the training time of the models. We also took advantage of the “super-convergence” phenomenon by using the one learning rate cycle strategy [34]. The detection task was approached as a regression problem to map a sequence to a sleepiness index. Therefore, we employed the Adam optimizer [35] with the mean absolute error as loss function. KSS scoring scale was used as ground-truth.

In test, five different models\(^2\) initialization were trained on the development and training sets together. The training batch size was reduced at eight to increase the generalization capability of our system. Then, an ensembling strategy was applied by averaging the final score of the five individuals models previously trained. Table 3 show our results in development and test. These results are compared with a set of relevant results published on the Continuous Sleepiness Estimation corpus.

Table 3: Results on development and test. Comparison with previously published results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2S+(13MFCC+(\Delta + \Delta\Delta))-Energy-ZC</td>
<td>.257</td>
<td>.364</td>
</tr>
<tr>
<td>ComParE-SVM [12]</td>
<td>.127</td>
<td>.172</td>
</tr>
<tr>
<td>ComParE-BoAW-SVM [12]</td>
<td>.250</td>
<td>.304</td>
</tr>
<tr>
<td>AUDEEP: S2SAE + SVM [12]</td>
<td>.243</td>
<td>.325</td>
</tr>
<tr>
<td>Fusion-I [12]</td>
<td>-</td>
<td>.343</td>
</tr>
<tr>
<td>ComParE + FV + BoAW [36]</td>
<td>.367</td>
<td>.383</td>
</tr>
<tr>
<td>ComParE16-MFCC-VQual [37]</td>
<td>.300</td>
<td>.331</td>
</tr>
<tr>
<td>Ordinal Triplet Loss &amp; BoAW-2000 [38]</td>
<td>.343</td>
<td>-</td>
</tr>
</tbody>
</table>

One of the benefits of our system is the use of low-dimensional feature vectors (42 dimensions) to accomplish the detection task. Most of the system previously published work on high-dimensional functional representation of Low Level Descriptors (LLD), such as the ComParE feature representation (6,373 dimensions) [39]. Therefore, we use a more compact representation of the information which requires less computational resources and reduces curse of dimensionality risks.

In development, our results are over the baseline, however, get outperformed by systems with the combination of LLD, prosody and/or their statistical modelling. Nevertheless, we get the second best result in testing. We attribute this behaviour to three main points:

- Ensembling score strategy to reduce variance on results.
- Small batch size to improve generalization capacity of the model [40]

\(^2\)As the results of 5 and 10 models ensembling were quite similar in development, we decided to use just 5 models ensembling in test to reducing complexity.
• The number of available training samples increases remarkably for testing. In front of a high volume of data, complex DNN models (e.g., our S2S proposal) usually have a better performance than classical machine learning classifiers (e.g., Support Vector Machine, the most used in previous works)[41, 42].

5. Conclusions
A Sequence-to-Sequence (S2S) model with global attention mechanism is proposed as sleepiness detection system. To the best of the author’s knowledge, this is the first of its kind in being proposed for this task. Speech patterns related to the sleepiness state are argued. The SLEEP (Düsseldorf Sleepy Language) Corpus, from the INTERSPEECH 2019 Computational Paralinguistics Challenge, is analysed as experimental framework. Our results are compared with the most relevant systems previously published. During development, we achieved a performance ($\rho = 0.257$) over the baseline system results in 2019. A score ensembling strategy is applied to deliver the final decision. Then, with a remarkable increment of training samples for test, our system performance was boosted to a $\rho$ of 0.364, being the second best system. This behaviour is because of the relationship between model complexity, database size and performance. It shows the need and benefits of increasing the study of deep learning approaches in the task of sleepiness state detection.

6. Acknowledgements
This work has received financial support from Axudas propias para a mobilidade de Personal Investigador da Universidade de Vigo 2021, the Xunta de Galicia (Centro singular para a mobilidade de Persoal Investigador da Unversidade de Vigo 2021, the Xunta de Galicia (Centro singular de investigación y transferencia de conocimiento de Galicia accreditation 2019-2022), Consellería de Cultura (Educación e Ordenación Universitaria; axudas para a consolidación e estruturación de unidades de investigación competitivas do Sistema Universitario de Galicia -ED431B 2021/24), and the European Union (European Regional Development Fund - ERDF). Moreover, the contributions of Prof. Krajewski and Wittenborn have been supported by the German Federal Ministry of Education and Research (BMBF) Project TRAM (FKZ 13FH005 SB7).

7. References
[22] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratoiu, R. Cowie, and


